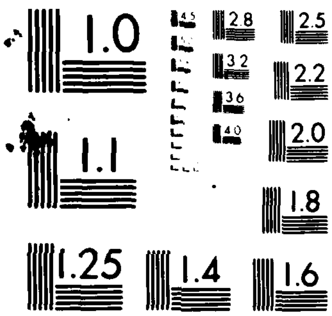
AD-A107 109 MASSACHUSETTS UNIV AMHERST DEPT OF MATHEMATICS AND S--ETC F/G 12/1
RANK CORRELATION COEFFICIENTS AS OPTIMALITY MEASURES FOR ORDINA--ETC(U)
MAY 81 M F JANOSWITZ N00014-79-C-0629
UNCLASSIFIED TR-J8101 NL

END
DATE
FILMED
12 81
DTIC



MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS-1963-A

AD A107109

LEVEL II

(12)

Technical Report J8101

RANK CORRELATION COEFFICIENTS AS OPTIMALITY MEASURES FOR ORDINAL CLUSTER METHODS

by

M. F. Janowitz

Department of Mathematics and Statistics
University of Massachusetts
Amherst, Massachusetts 01003

May, 1981

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE — READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: J8101	2. GOVT ACCESSION NO.: AD-A107 109
	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): Rank Correlation coefficients as optimality measures for ordinal cluster methods.	5. TYPE OF REPORT & PERIOD COVERED: Technical
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): M. F. Janowitz	8. CONTRACT OR GRANT NUMBER(s): N-00014-79-C-0629
9. PERFORMING ORGANIZATION NAME AND ADDRESS: University of Massachusetts, Amherst, MA 01003	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 121405
11. CONTROLLING OFFICE NAME AND ADDRESS: Procuring Contracting Officer, Office of Naval Research, Arlington, VA 22217	12. REPORT DATE: April 1981
	13. NUMBER OF PAGES
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Office of Naval Research Resident Representative, Harvard University, Gordon McKay Laboratory, Room 113, Cambridge, MA 02138	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Cluster analysis, dissimilarity coefficient, optimality measure, rank order correlation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Use and misuse of rank correlation coefficients as optimality measures for cluster analysis is investigated. Distributions of some of the more common dissimilarity coefficients on random binary data are established. The distributions of certain rank correlation coefficients between input and output of cluster methods is also investigated for random input data. A statistical model is proposed to deal with this type of question.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

RANK CORRELATION COEFFICIENTS AS OPTIMALITY MEASURES FOR ORDINAL CLUSTER METHODS

by M. F. Janowitz

§0. Introduction. The reader who is not familiar wit cluster analysis is referred to [11] for an introduction to the subject. The input data in a clustering problem often consists of a finite set P of objects to be classified and a finite set A of attributes that the objects of P might possess. For purposes of this investigation, it will always be assumed that the attributes are binary in that an element of P either has or does not have a given attribute. If there are p objects in P, and n attributes, then the input data can be thought of as a $p \times n$ matrix $A = (a_{ij})$, where a_{ij} is 1 or 0 according to whether the ith object has or does not have attribute j. Cluster analysis frequently takes the form of a 2 step process:

Step 1. The attribute data is converted to a numerical measure of dissimilarity called a dissimilarity coefficient. This is simply a mapping d from $P \times P$ to the nonnegative reals such that: (i) $d(x,y) = d(y,x)$, and (ii) $d(x,y) = 0$ if and only if $x = y$, with these conditions holding for every x,y in P.

Step. 2. The dissimilarity coefficient d is converted to a hierarchical stratified clustering. This is a sequence $E_1 \subseteq E_2 \subseteq \ldots \subseteq E_t = P \times P$ of equivalence relations on P with E_i coming into being at level h_i, where $h_1 \le h_2 \le \ldots \le h_t$ is a sequence of nonnegative real numbers. Many commonly used cluster methods are ordinal in that they only consider the ranks of the input dissimilarity coefficient. For such cluster methods, one often views the output as an ordinal dissimilarity coefficient d' rather than as a stratified clustering. The idea is to define $d'(x,y)$ to be the smallest positive integer for which xE_iy.

The problem now that faces the investigator is the determination of how well the output clustering fits his input data. The usual procedure is to compare the output d' with the intermediate dissimilarity coefficient d, and hope that if they match well, then the output will provide an accurate reflection of the original attribute data. There is of course some danger involved in making this assumption, as was shown in [7]. A number of possible optimality measures for this type of clustering were considered in [8]. Since the intermediate and output dissimilarity coefficients are each assumed to be ordinal, I shall restrict my attention here to the consideration of the two most commonly used rank order correlation coefficients: Kendall's tau-coefficient, and Spearman's rho-coefficient. They are defined as in Kendall [9] with corrections for ties as given by [9], (3.3), p.35 and [9], (3.8), p. 38.

There are now at least 3 questions that can be asked, and I shall consider them separately.

First Question. Suppose two cluster methods are applied to the same dissimilarity measure d. Can rho or tau be used to determine

Presented in part to the Classification Society
June 1, 1981.

which output provides a better fit to d? Schematically, we have the following situation:

$$A \longrightarrow d \begin{array}{l} \nearrow d_1' \\ \searrow d_2' \end{array}$$

<u>Second Question</u>. Given a fixed intermediate dissimilarity coefficient d and a fixed output d', can <u>rho</u> or <u>tau</u> be used to decide whether d' came about by chance or whether it actually reflects some structure in P?

<u>Third Question</u>. For a fixed attribute matrix A, suppose d_1 and d_2 are competing intermediate dissimilarity coefficients, and that cluster methods are applied to d_1 and d_2 to produce outputs d_1' and d_2'. Can <u>rho</u> or <u>tau</u> applied to the pairs (d_1,d_1'), (d_2,d_2') be used to determine which of d_1' or d_2' provides the better fit to the original attribute data A?

The above questions will be dealt with in succession in the next three sections of the paper. Following this, there will be four sections dealing with the distributions of certain dissimilarity measures on binary random attribute data.

To avoid needless complications, we shall focus our attention on only one cluster method: <u>single linkage clustering.</u> In much the same spirit, it will be useful to limit the possible dissimilarity coefficients to four choices: the <u>simple matching coefficent</u>, the <u>coefficient of Russell and Rao</u>, <u>Jaccard's coefficient</u>, and the <u>coefficient of special similarity</u>. To see how these are defined, consider n binary attributes on the

elements x,y of P. Let a be the number of common 1's, c the number of common 0's, and b the number of mismatches. The first three coefficients are then respectively defined by subtracting from 1 the quantities $(a+c)/n$, a/n and $a/(a+b)$. Special similarity will not be defined until §7.

§1. <u>The first question</u>. Recall that we are given a fixed intermediate dissimilarity coefficient d, and two competing output coefficients d_1' and d_2'. We are to determine which of d_1' or d_2' provides the better fit to d.

This question is very easy to answer. Notice first that we are not dealing with a sample from a larger population. The values of d are fixed and represent all of the data under consideration. Thus we need not be concerned with questions of statistical significance. In this situation, both <u>tau</u> and <u>rho</u> seem to provide useful measures of goodness of fit. To see this, we need only consider how these coefficients are defined. Each of them is represented by a fraction whose denominator is a normalizing factor. To see how the numerators are defined, let us imagine that we are trying to compare the sequences of values $(a_1,a_2,\ldots,a_k)$ and $(b_1,b_2,\ldots,b_k)$ where each sequence is rank ordered as described in Kendall ([9], p.34). Thus the values (.5,8,10,8) would be represented as (1,2.5,4,2.5), etc. For $i < j \le k$, we now define $d_{ij} = 1$, 0 or -1 according to whether $(a_i - a_j)(b_i - b_j)$ is positive, zero or negative. The numerator of Kendall's coefficient is then simply the sum of the d_{ij}'s, while that of Spearman's coefficient

is the sum of the numbers $(a_i - a_j)(b_i - b_j)$. Questions of statistical significance sometimes arise, and are dealt with in connection with the "second question" (See §2).

Even for ordinal cluster methods, the use of a rank order correlation coefficient can cause some problems. This was mentioned briefly in [8], and will again be illustrated here. The problem is simply that a rank order coefficient rank orders things, and this may or may not be desirable. For the input data d on the set $\{w,x,y,z\}$ each of the outputs d_i' $(i = 2,3,4,5,6)$ would be treated by rho and tau as providing the same goodness of fit to d, whereas one might argue that, for example, $i = 4$ might be superior to either $i = 2$ or $i = 6$.

	wx	wy	wz	xy	xz	yz
d	1	2	3	4	5	6
d'	1	i	i	i	i	i

§2. The second question. Given a fixed intermediate dissimilarity coefficient d and a fixed output d', can rho or tau be used to decide whether d' came about by chance or whether it actually reflects the structure of P?

The answer is yes, but the process must be approached with some caution. The problem involves the formulation of a suitable null hypothesis. This was noted by Hubert and Baker [5], and the present paper should be considered as an extension of their work. By means of some computer simulations, it will be shown that for each of the three

choices of dissimilarity coefficient, one must assume the possibility of ties when one is dealing with random attribute data. It will also be shown that each of these coefficients imposes some structure on random data. For these reasons one cannot take as a null hypothesis the assertion that the values of the intermediate dissimilarity coefficient are rank ordered with no ties and with each possible ranking equally likely. For purposes of future reference this assumption will be referred to as the permutation model. A possible choice of null hypothesis is

H_0: The input attribute data consists of n random binary attributes with probability 0.5 of a I occurring.

This will be referred to as Model n. If H_0 is true, we shall see that *neither tau nor rho has anything like a normal distribution, and that* their distributions depend on the choice of dissimilarity coefficient, the number of attributes, and of course the number of objects in the set to be classified.

Goodall [5] considers a model in which attribute i is random with probability p_i of a I occurring.

Before proceeding, a few words are in order regarding notation. Each choice of dissimilarity coefficient was applied to random binary attribute data defined on a 4 element, a 5 element, and a 6 element set. This was followed by single linkage clustering and the computation of both tau and rho. Each simulation involved 500 trials. If the single linkage output involved no observable structure, the corresponding case was discarded, as neither tau nor rho could be defined. To investigate

the effect of the number of attributes, separate simulations were carried out for 10, 25 and 50 attributes. So that meaningful pairwise comparisions could be made, each trial involved computing the 3 dissimilarity coefficients on the same input data. A consistent notation was devised, and here it is:

KENS25 represents the values of the Kendall coefficient using simple matching and 25 attributes. The corresponding values for the coefficient of Russell & Rao and that of Jaccard are denoted KENR25 and KENJ25. To denote the Spearman coefficient, KEN is replaced by SP. Finally, KPERM and SPERM represent values of the Kendall and Spearman coefficients for the permutation model.

Descriptive statistics for the Kendall and Spearman coefficients appear in Tables 1, 2 and 3. An examination of these statistics now leads to several conclusions. Each choice of dissimilarity coefficient produces values of tau and rho that are inflated over those that would be expected in the permutation model, with the coefficient of Russell & Rao producing the most inflation, Jaccard usually second, and simple matching third. It would seem from this that the simple matching coefficient imposes less structure on random attribute data than either of the other choices. Secondly, the distributions of both tau and rho for random attribute data possess a much higher degree of negative skewness than one would expect under the permutation model. Thirdly, as the number of attributes goes up, the means of both rho and tau seem to decrease. This is probably caused by the presence of fewer ties in the

intermediate dissimilarity coefficient. Finally, Russell & Rao seems most likely to produce a tree reflecting no structure, with simple matching second and Jaccard third.

Next we turn to the question of normality of the various distributions. The high degree of skewness exhibited in Tables 1, 2 and 3 already make it rather unlikely that any of the distributions should be even close to normal, as does the display of histograms for Model 50 shown in the Appendix. Kolmogoroff-Smirnoff tests for normality were also performed, and here are the results for Model 50.

DESCRIPTIVE STATISTICS

VARIABLE	N	MIN.	MAX.	RANGE	MEAN	VARIANCE	ST. DEV.	ST. ERR.	SKEWNESS	KURTOSIS
KPERM	5	0.3890	0.8560	0.4670	0.6532	0.0349	0.1869	0.0836	¯0.2894	¯1.8653
KENS10	468	0.2770	1.0000	0.7230	0.7275	0.0241	0.1553	0.0072	¯0.5859	0.1998
KENS25	478	0.2770	1.0000	0.7230	0.7122	0.0224	0.1495	0.0068	¯0.6684	0.0090
KENS50	495	0.2670	1.0000	0.7330	0.6792	0.0276	0.1660	0.0075	¯0.5997	¯0.4660
KENF10	444	0.2770	1.0000	0.7230	0.7615	0.0231	0.1519	0.0072	¯0.6011	0.5066
KENF25	469	0.2670	1.0000	0.7330	0.7277	0.0221	0.1486	0.0069	¯0.5479	¯0.0190
KENF50	481	0.2670	1.0000	0.7330	0.7168	0.0237	0.1541	0.0070	¯0.6297	¯0.2818
KENJ10	493	0.2670	1.0000	0.7330	0.7215	0.0258	0.1606	0.0072	¯0.7424	¯0.3315
KENJ25	500	0.2670	1.0000	0.7330	0.6992	0.0246	0.1569	0.0070	¯0.7048	¯0.5314
KENJ50	499	0.2670	0.9200	0.6530	0.6871	0.0276	0.1661	0.0074	¯0.6533	¯0.7967

THERE ARE 4 OBJECTS TO BE CLASSIFIED

++

DESCRIPTIVE STATISTICS

VARIABLE	N	MIN.	MAX.	RANGE	MEAN	VARIANCE	ST. DEV.	ST. ERR.	SKEWNESS	KURTOSIS
SPERM	5	0.4630	0.9260	0.4630	0.7246	0.0343	0.1853	0.0829	¯0.2870	¯1.8667
SPS10	468	0.3020	1.0000	0.6980	0.7741	0.0236	0.1535	0.0071	¯0.8220	0.5709
SPS25	478	0.3020	1.0000	0.6980	0.7669	0.0223	0.1495	0.0068	¯0.8456	0.3592
SPS50	495	0.2970	1.0000	0.7030	0.7389	0.0279	0.1670	0.0075	¯0.7136	¯0.2065
SPF10	444	0.3020	1.0000	0.6980	0.7989	0.0224	0.1495	0.0071	¯0.8345	0.8586
SPF25	469	0.2970	1.0000	0.7030	0.7766	0.0216	0.1470	0.0068	¯0.6560	0.1906
SPF50	481	0.2970	1.0000	0.7030	0.7712	0.0235	0.1532	0.0070	¯0.7227	¯0.0656
SPJ10	493	0.2970	1.0000	0.7030	0.7814	0.0253	0.1592	0.0072	¯0.8643	¯0.1008
SPJ25	500	0.2970	1.0000	0.7030	0.7655	0.0242	0.1557	0.0070	¯0.7408	¯0.4288
SPJ50	499	0.2970	0.9530	0.6560	0.7554	0.0272	0.1650	0.0074	¯0.6838	¯0.7063

THERE ARE 4 OBJECTS TO BE CLASSIFIED

Table 1.

DESCRIPTIVE STATISTICS

VARIABLE	N	MIN.	MAX.	RANGE	MEAN	VARIANCE	ST. DEV.	ST. ERR.	SKEWNESS	KURTOSIS
KPERM	500	0.1760	0.8820	0.7060	0.5403	0.0230	0.1518	0.0068	¯0.1546	¯0.5882
KENS10	484	0.1320	0.9440	0.8120	0.6410	0.0238	0.1544	0.0070	¯0.5013	¯0.3155
KENS25	491	0.1330	0.9240	0.7910	0.6075	0.0250	0.1582	0.0071	¯0.4092	¯0.3091
KENS50	500	0.1320	0.9130	0.7810	0.6039	0.0224	0.1497	0.0067	¯0.3676	¯0.3011
KENF10	477	0.2720	1.0000	0.7280	0.7001	0.0218	0.1476	0.0068	¯0.4611	¯0.3870
KENF25	494	0.1570	0.9600	0.8030	0.6527	0.0252	0.1588	0.0071	¯0.3520	¯0.5687
KENF50	495	0.1510	0.9320	0.7810	0.6547	0.0231	0.1521	0.0068	¯0.5465	¯0.2099
KENJ10	498	0.1070	0.9600	0.8530	0.6449	0.0261	0.1615	0.0072	¯0.6894	0.2132
KENJ25	500	0.1290	0.9350	0.8060	0.6188	0.0245	0.1565	0.0070	¯0.3118	¯0.5981
KENJ50	500	0.1260	0.8820	0.7560	0.6140	0.0244	0.1563	0.0070	¯0.5842	¯0.1268

THERE ARE 5 OBJECTS TO BE CLASSIFIED

DESCRIPTIVE STATISTICS

VARIABLE	N	MIN.	MAX.	RANGE	MEAN	VARIANCE	ST. DEV.	ST. ERR.	SKEWNESS	KURTOSIS
SPERM	500	0.2290	0.9530	0.7240	0.6210	0.0265	0.1629	0.0073	¯0.2394	¯0.7104
SPS10	484	0.1460	0.9740	0.8280	0.6946	0.0259	0.1611	0.0073	¯0.5848	¯0.2311
SPS25	491	0.1810	0.9650	0.7840	0.6718	0.0276	0.1661	0.0075	¯0.5146	¯0.2708
SPS50	500	0.1460	0.9650	0.8190	0.6723	0.0247	0.1572	0.0070	¯0.4429	¯0.3144
SPF10	477	0.2950	1.0000	0.7050	0.7448	0.0229	0.1512	0.0069	¯0.5673	¯0.3768
SPF25	494	0.1900	0.9870	0.7970	0.7119	0.0270	0.1645	0.0074	¯0.4630	¯0.5214
SPF50	495	0.1870	0.9710	0.7840	0.7224	0.0253	0.1591	0.0072	¯0.6833	¯0.0981
SPJ10	498	0.1130	0.9870	0.8740	0.7170	0.0283	0.1683	0.0075	¯0.8517	0.4429
SPJ25	500	0.1980	0.9710	0.7730	0.7001	0.0264	0.1626	0.0073	¯0.4322	¯0.5810
SPJ50	500	0.1970	0.9530	0.7560	0.6980	0.0268	0.1638	0.0073	¯0.7106	¯0.0584

THERE ARE 5 OBJECTS TO BE CLASSIFIED

Table 2.

DESCRIPT'VE STATISTICS

VARIABLE	N	MIN.	MAX.	RANGE	MEAN	VARIANCE	ST. DEV.	ST. ERR.	SKEWNESS	KURTOSIS
KPERM	500	0.1160	0.7880	0.6720	0.4522	0.0198	0.1408	0.0063	0.0427	¯0.5946
KENS10	493	0.1120	0.9580	0.8460	0.5803	0.0228	0.1510	0.0068	¯0.4524	¯0.0160
KENS25	499	0.0580	0.8820	0.8240	0.5458	0.0238	0.1544	0.0069	¯0.4382	¯0.1675
KENS50	499	0.0750	0.8620	0.7870	0.5233	0.0223	0.1493	0.0067	¯0.3897	¯0.1408
KENR10	488	0.0630	0.9480	0.8850	0.6555	0.0226	0.1503	0.0068	¯0.6111	¯0.0343
KENR25	492	0.0620	0.9100	0.8480	0.6331	0.0213	0.1459	0.0066	¯0.5803	0.1544
KENR50	499	0.1550	0.9300	0.7750	0.6212	0.0204	0.1427	0.0064	¯0.4283	¯0.1546
KENJ10	500	0.0900	0.9110	0.8210	0.5978	0.0200	0.1414	0.0063	¯0.4449	¯0.0188
KENJ25	500	0.0760	0.8830	0.8070	0.5653	0.0203	0.1424	0.0064	¯0.5136	0.1996
KENJ50	500	0.1040	0.8790	0.7750	0.5491	0.0216	0.1468	0.0066	¯0.2369	¯0.3643

THERE ARE 6 OBJECTS TO BE CLASSIFIED

DESCRIPTIVE STATISTICS

VARIABLE	N	MIN.	MAX.	RANGE	MEAN	VARIANCE	ST. DEV.	ST. ERR.	SKEWNESS	KURTOSIS
SPERM	500	0.1660	0.8980	0.7320	0.5365	0.0250	0.1580	0.0071	¯0.0208	¯0.6544
SPS10	493	0.1340	0.9840	0.8500	0.6373	0.0260	0.1613	0.0073	¯0.5341	¯0.0122
SPS25	499	0.0660	0.9470	0.8810	0.6136	0.0277	0.1664	0.0074	¯0.5361	¯0.0910
SPS50	499	0.0980	0.9340	0.8360	0.5958	0.0258	0.1605	0.0072	¯0.5014	0.0097
SPR10	488	0.0690	0.9820	0.9130	0.7044	0.0253	0.1589	0.0072	¯0.7080	0.0265
SPR25	492	0.0680	0.9500	0.8820	0.6997	0.0240	0.1549	0.0070	¯0.7286	0.3623
SPR50	499	0.1810	0.9750	0.7940	0.7002	0.0232	0.1524	0.0068	¯0.6211	¯0.0061
SPJ10	500	0.1130	0.9670	0.8540	0.6807	0.0229	0.1513	0.0068	¯0.6385	0.1569
SPJ25	500	0.1230	0.9560	0.8330	0.6595	0.0238	0.1543	0.0069	¯0.6894	0.3532
SPJ50	500	0.1270	0.9590	0.8320	0.6433	0.0251	0.1585	0.0071	¯0.4331	¯0.2170

THERE ARE 6 OBJECTS TO BE CLASSIFIED

Table 3.

Objects	Variable	Reject Normality	P-value	Variable	Reject Normality	P-value
4	KENS 50	Yes	.01	SPS 50	Yes	.01
	KENR 50	Yes	.01	SPR 50	Yes	.01
	KENJ 50	Yes	.01	SPJ 50	Yes	.01
5	KPERM	Yes	.01	SPERM	Yes	.01
	KENS 50	Yes	.01	SPS 50	Yes	.01
	KENR 50	Yes	.01	SPR 50	Yes	.01
	KENJ 50	Yes	.01	SPJ 50	Yes	.01
6	KPERM	Yes	.10	SPERM	No	
	KENS 50	Yes	.025	SPS 50	Yes	.01
	KENR 50	Yes	.01	SPR 50	Yes	.01
	KENJ 50	Yes	.10	SPJ 50	Yes	.01

The normality of the distribution of KPERM and SPERM on a 4 element set was not considered because these distributions were completely determined. There are only 5 possible values for either the tau or the rho coefficient and they are equally likely to occur. If $P = \{w,x,y,z\}$. suppose wx has rank 1. The 5 possibilities are either that yz have rank 2, or that say wy have rank 2 and xy have respectively rank 3, 4, 5 or 6. The values of tau and rho are then:

tau	rho	Explanation
0.775	0.845	yz has rank 2
0.856	0.926	wy rank 2, xy rank 3
0.701	0.772	wy rank 2, xy rank 4
.0545	0.617	wy rank 2, xy rank 5
0.389	0.463	wy rank 2, xy rank 6

This leaves us with the question of whether one can safely ignore the possibility of ties in the intermediate dissimilarity coefficient. To check this, an easily computable numerical measure of the number of ties was devised. The intermediate dissimilarity coefficient was rank ordered with ties having the same rank. The difference between the highest possible rank with no ties, and the actual highest rank was then used as a measure of the number of ties. This can best be illustrated by considering an example or two:

Ranking	Highest Poss.	Actual Highest	Difference
1 2 3 5 4 6	6	6	0
1 2 3 2 4 5	6	5	1
1 2 3 2 3 4	6	4	2
1 2 3 2 2 4	6	4	2
1 2 2 2 2 3	6	3	3

The figures in the last column represent the measure of tieing. With sets of 4,5,6 elements, and with 10,25,50 attributes, 500 trials were performed using random binary attribute data and each of the three dissimilarity coefficients that have been considered. The results appear in Table 4. The columns labeled Expected No. Ties represent in each case the average of the measure we just defined. A glance at the table should convince one that it is very dangerous to ignore the possibility of ties - even for Jaccard's coefficient, which is the one that is least likely to produce a tie.

		Simple Matching		Russell and Rao		Jaccard	
No. Elts.	No. Char.	Prob. 0 ties	Expected No. Ties	Prob. 0 ties	Expected No. Ties	Prob. 0 ties	Expected No. Ties
4	10	.03	2.03	.002	2.63	.31	1.06
	25	.13	1.46	.05	1.85	.68	0.40
	50	.28	1.04	.20	1.31	.82	0.20
5	10	.002	4.97	.004	5.79	.03	2.69
	25	.02	3.66	.02	4.45	.34	1.05
	50	.002	2.85	.04	3.50	.58	1.13
6	10	0	9.09	0	10.24	0	1.05
	25	0	7.22	0	8.33	.10	2.25
	50	0	5.87	0	6.98	.30	1.24

Table 4. Summary of results on ties

§3. The third question. Here d_1 and d_2 are distinct dissimilarity measures, A is a fixed attribute matrix, and d_1', d_2' are outputs from applying possibly different cluster methods to d_1 and d_2. Suppose rho or tau is higher for the pair d_1, d_1' than it is for the pair d_2, d_2'. Can one conclude that somehow d_1' is more likely than d_2' to reflect the actual structure of the underlying set P?

Evidence from computer simulations shows that this is an extremely dangerous type of conclusion to draw. The reason is simple. Different dissimilarity coefficients treat random data in different ways. Thus the high value of the correlation coefficient for the pair d_1, d_1' could be largely due to chance error, while the pair d_2, d_2' largely ignores that error. To see this, one need only consider Table 5, where product moment

correlations are presented for the values of rho and tau on identical random data.

No. Elts	Char.	Kendall SR	Kendall SJ	Kendall RJ	Spearman SR	Spearman SJ	Spearman RJ
4	10	.2583	.4799	.6287	.2541	.4698	.6457
	25	.2121	.4738	.5944	.1956	.4664	.5958
	50	.0880	.3626	.5744	.0882	.3605	.5842
5	10	.0243	.3727	.5541	.0262	.3720	.5636
	25	.1707	.3784	.6262	.1545	.3768	.6272
	50	.1390	.3252	.6186	.1248	.3243	.6263
6	10	.1153	.3255	.6139	.1168	.3204	.5974
	25	.1466	.3295	.6410	.1303	.3253	.6512
	50	.0931	.2884	.6432	.0728	.2667	.6511

Table 5. Product moment correlations for Kendall and Spearman coefficients between differing dissimilarity coefficients on random binary data. Each figure represents 500 trials with those cases discarded for which the correlations are not defined. The columns labeled SR denote the comparison between simple matching and Russell & Rao, SJ those between simple matching and Jaccard, and RJ those of Russell & Rao with Jaccard.

It is also pertinent to consult Janowitz [7] where a similar question is considered using the product moment correlation as a measure of optimality.

§4. Distribution for the simple matching coefficient.

Here we shall try to see just why the simple matching coefficient seems to impose structure on random data. Suppose that we are given n binary attributes on the set P. For fixed elements x and y, it will be convenient to consider the values of $SS(x,y)$ in place of those of the simple matching coefficient, where SS denotes the number of attributes which either both x and y possess or which neither of them possesses. Thus the value of the simple matching coefficient would be $1 - SS(x,y)/n$. An examination of the possible values of attributes on x,y

x	y	SS(x,y)
1	1	1
1	0	0
0	1	0
0	0	1

shows that on random data, the values of SS follow a binomial distribution with probability 0.5 for SS to occur. This means that for $i \leq n$, the probability that $SS(x,y) = i$ is simply $(n;i) \times (.5)^n$, where $(n;i)$ denotes the binomial coefficient $n!/(i!)(n-i)!$. It follows from this that the expected value of $SS(x,y)$ is $n/2$, and that of the simple matching coefficient 0.5.

Let us now put a third element z into the picture and examine the possible attributes on the triple $\{x,y,z\}$.

x	1	1	0	0	1	1	0	0
y	1	1	0	0	0	0	1	1
z	1	0	1	0	1	0	1	0
SS(x,z)	1	0	0	1	1	0	0	1

It is clear from this that the distribution of values for $SS(x,z)$ is independent from that of $SS(x,y)$. Similarly, the values of $SS(z,w)$ are independent from those of $SS(x,y)$. Thus any two values of SS are independent. Despite this, as we shall soon see, the values of SS on a triple of elements need not be independent.

Theorem 1. Let $SS(x,y) = k$ and $SS(x,z) = j$ with $j \le k$. Then

$$j + (n - k) \ge SS(y,z) \ge |(j + k) - n|.$$

Proof: Perhaps the easiest way to establish this result is to simply consider the possibilities. Because the simple matching coefficient treats the attribute states 0 and 1 symmetrically, we first observe that we need only consider 4 possible attributes on $\{x,y,z\}$ as follows:

Number of Attributes						
	x	1	1	1	1	
	y	1	1	0	0	
	z	1	0	0	1	SS(y,z)
		j	k-j	n-k	0	j + (n - k)
		j-1	(k-j)+1	n-k-1	1	j + (n - k) - 2
		j-i	(k-j)+i	n-k-i	i	j + (n - k) - 2i
	Case 1	j-(n-k)	n-j	0	n-k	(j + k) - n
	Case 2	0	k	n-k-j	j	n - (j + k)

Case 1. $j + k \ge n$ Case 2. $j + k < n$

Corollary. $SS(y,z)$ can take on at most only min $[1+(n-k), 1+k]$ where consecutive values differ by 2.

It should be noted that most of Theorem 1 can be established from the well known fact that the simple matching coefficient is a metric. The only new item is the fact that if $j + k < n$, then

$$n - (j + k) \le SS(y,z).$$

This says that for arbitrary values of j,k one has

$$n \le SS(x,y) + SS(x,z) + SS(y,z)$$

from which it follows that

$$2n \ge [n - SS(x,y)] + [n - SS(y,z)] + [n - SS(x,z)].$$

Letting d_S denote the value of the simple matching coefficient, this establishes

Theorem 2. The simple matching coefficient d_S is a metric taking values in the interval $[0,1]$ and having the further property that for

all elements x,y,z of P,

$$d(x,y) + d(x,z) + d(y,z) \le 2.$$

This expresses the rather bizarre property that if both y and z are far from an element x, they must be close to each other!

We turn now to the distribution of $SS(y,z)$ on random data, when the values of $SS(x,y)$ and $SS(x,z)$ are known. As shown in the proof of Theorem 1,, there are only 4 attributes that need be considered and they are equally likely to occur. Given nonnegative integers $i_1,i_2,\ldots,i_s$ whose sum is n, it will be convenient to let $(n;i_1,i_2,\ldots,i_s)$ denote the multinomial coefficient $n!/(i_1!)(i_2!)\ldots(i_s!)$. For fixed values of j and k, the probability of obtaining $SS(y,z) = j + (n-k) - 2i$ $(0 \le i \le \min\{j,n-k\})$ is $(n;j-i,(k-j)+i,n-k-i,i)$ divided by $\Sigma_t(n;j-t,(k-j)+t,n-k-t,t)$, where t goes from 0 to $\min\{j,n-k\}$. Table 6 contains the distribution of $SS(y,z)$ for 5 attributes and Table 7 for 8 attributes. In these tables, the first column represents the value of $SS(x,z)$, the second column that of $SS(x,y)$, the next to the last column the expected value of $SS(y,z)$, and the last column the probability that the value of $SS(y,z)$ will dominate or equal the value of $SS(v,w)$ for elements v,w distinct from x,y,z. The remaining columns represent the distribution of $SS(y,z)$ for the indicated values of $SS(x,y)$ and $SS(x,z)$. Thus in Table 6, to find the probability that $SS(y,z) = 2$, given that $SS(x,z) = 2$ and $SS(x,y) = 3$, one looks in the row that starts with the entries 2 3, and notes that under the column labeled 2.00, the probability is 0.60. To illustrate the computation of

this probability, we note that the possible values of $SS(y,z)$ are

x	1	1	1	1	
y	1	1	0	0	
z	1	0	0	1	SS(y,z)
	2	1	2	0	4
	1	2	1	1	
	0	3	0	2	0

The desired probability is then given by

$$(5;1,2,1,1)/[(5;2,1,2,0)+(5;1,2,1,1)+(5;0,3,0,2)] =$$
$$60/(30 + 60 + 10) = 0.60.$$

The probabilities shown in the last column of these tables should be compared with the *a priori* probabilities that the value of $SS(x,y)$ should dominate or equal that of $SS(v,w)$ where v,w are distinct from x,y. For 5 and 8 attributes, these are respectively 0.62 and 0.60.

			0.00	1.00	2.00	3.00	4.00	5.00		
0	5		1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03
1	5		0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.19
2	5		0.00	0.00	1.00	0.00	0.00	0.00	2.00	0.50
3	5		0.00	0.00	0.00	1.00	0.00	0.00	3.00	0.81
4	5		0.00	0.00	0.00	0.00	1.00	0.00	4.00	0.97
0	4		0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.19
1	4		0.20	0.00	0.80	0.00	0.00	0.00	1.60	0.41
2	4		0.00	0.40	0.00	0.60	0.00	0.00	2.20	0.56
3	4		0.00	0.00	0.60	0.00	0.40	0.00	2.80	0.69
0	3		0.00	0.00	1.00	0.00	0.00	0.00	2.00	0.50
1	3		0.00	0.40	0.00	0.60	0.00	0.00	2.20	0.56
2	3		0.10	0.00	0.60	0.00	0.30	0.00	2.40	0.59
0	2		0.00	0.00	0.00	1.00	0.00	0.00	3.00	0.81
1	2		0.00	0.00	0.60	0.00	0.40	0.00	2.80	0.69
0	1		0.00	0.00	0.00	0.00	1.00	0.00	4.00	0.97

Table 6. Conditional distributions for SS(y,z) for given values of SS(x,y) and SS(x,z). See text for explanation. These figures are for 5 attributes.

			0.00	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00		
0	8		1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	8		0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.04
2	8		0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00	0.14
3	8		0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	3.00	0.36
4	8		0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	4.00	0.64
5	8		0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	5.00	0.86
6	8		0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	6.00	0.96
7	8		0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	7.00	1.00
0	7		0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.04
1	7		0.12	0.00	0.88	0.00	0.00	0.00	0.00	0.00	0.00	1.75	0.13
2	7		0.00	0.25	0.00	0.75	0.00	0.00	0.00	0.00	0.00	2.50	0.28
3	7		0.00	0.00	0.37	0.00	0.63	0.00	0.00	0.00	0.00	3.25	0.45
4	7		0.00	0.00	0.00	0.50	0.00	0.50	0.00	0.00	0.00	4.00	0.61
5	7		0.00	0.00	0.00	0.00	0.63	0.00	0.37	0.00	0.00	4.75	0.76
6	7		0.00	0.00	0.00	0.00	0.00	0.75	0.00	0.25	0.00	5.50	0.89
0	6		0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00	0.14
1	6		0.00	0.25	0.00	0.75	0.00	0.00	0.00	0.00	0.00	2.50	0.28
2	6		0.04	0.00	0.43	0.00	0.54	0.00	0.00	0.00	0.00	3.00	0.40
3	6		0.00	0.11	0.00	0.54	0.00	0.36	0.00	0.00	0.00	3.50	[illegible]
4	6		0.00	0.00	0.21	0.00	0.57	0.00	0.21	0.00	0.00	4.00	[illegible]
5	6		0.00	0.00	0.00	0.36	0.00	0.54	0.00	0.11	0.00	4.50	0.69
0	5		0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	3.00	[illegible]
1	5		0.00	0.00	0.37	0.00	0.63	0.00	0.00	0.00	0.00	3.25	[illegible]
2	5		0.00	0.11	0.00	0.54	0.00	0.36	0.00	0.00	0.00	3.50	[illegible]
3	5		0.02	0.00	0.27	0.00	0.54	0.00	0.18	0.00	0.00	3.75	0.55
4	5		0.00	0.07	0.00	0.43	0.00	0.43	0.00	0.07	0.00	4.00	0.60
0	4		0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	4.00	0.64
1	4		0.00	0.00	0.00	0.50	0.00	0.50	0.00	0.00	0.00	4.00	0.61
2	4		0.00	0.00	0.21	0.00	0.57	0.00	0.21	0.00	0.00	4.00	0.60
3	4		0.00	0.07	0.00	0.43	0.00	0.43	0.00	0.07	0.00	4.00	0.60
0	3		0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	5.00	0.86
1	3		0.00	0.00	0.00	0.00	0.63	0.00	0.37	0.00	0.00	4.75	0.76
2	3		0.00	0.00	0.00	0.36	0.00	0.54	0.00	0.11	0.00	4.50	0.69
0	2		0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	6.00	0.96
1	2		0.00	0.00	0.00	0.00	0.00	0.75	0.00	0.25	0.00	5.50	0.89
0	1		0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	7.00	1.00

Table 7. Conditional distributions for SS(y,z) for given values of SS(x,y) and SS(x,z). See text for explanation. These figures are for 8 attributes.

§5. Distribution of the coefficient of Russell & Rao.

Let us denote the value of this coefficient by d_R. Letting $A(x,y)$ represent the number of attributes shared by x and y, one then has that $d_R(x,y) = 1 - A(x,y)/n$, where n is the total number of attributes. Consideration of the possible values of attributes on the elements x and y shows that the values of A follow a binomial distribution with probability 0.25 for A to occur; hence the expected value for A is $n/4$. Unlike the situation with the simple matching coefficient, the distribution of $A(x,z)$ turns out to be dependent on the value of $A(x,y)$. To see this, consider the possible attributes on x,y,z:

	1	2	3	4	5	6	7	8
x	1	1	1	1	0	0	0	0
y	1	1	0	0	1	1	0	0
z	1	0	1	0	1	0	1	0

Attributes 1 and 2 are the ones that contribute to $A(x,y)$, and for these attributes there is a probability of 1/2 that there will also be a contribution to $A(x,z)$. For the remaining attributes, the probability drops to 1/6 for an $A(x,z)$ contribution. Thus if $A(x,y) = k$, one may compute the probability that $A(x,z) = j$ by viewing the process as the selection of k attributes from the first 2 columns, followed by an independent selection of $n-k$ attributes from among those remaining. The probability is then given by the sum

$$\sum_i (k;i)(1/2)^k(n-k;j-i)(1/6)^{j-i}(5/6)^{(n-k)-(j-i)}.$$

Specifically, if $n = 5$, $A(x,y) = 4$, then the probability that $A(x,z) = 2$ is given by

$$(4;2)(1/2)^4(1;0)(1/6)^0(5/6)^1 + (4;1)(1/2)^4(1;1)(1/6)^1(5/6)^0 =$$
$$(0.375)(0.833) + (0.25)(0.167) = 0.312 + .0042 = 0.354\ .$$

Table 8 follows the pattern of the earlier tables and gives the distribution of $A(x,z)$ for fixed values of $A(x,y)$ for 5 attributes. The left hand column denotes the value of $A(x,y)$ and the two right hand columns have the same meaning as they did in the earlier tables.

It is easy to show that d_R is a metric, and from this the next theorem is immediate. Alternately, one could proceed as in the proof of Theorem 1.

Theorem 3. *For* n *attributes, if* $A(x,y) = k$ *and* $A(x,z) = j$, *then*

$$(j + k) - n \le A(y,z) \le j + (n - k).$$

Naturally, this imposes further structure on the distribution of $A(y,z)$, given the values of $A(x,y)$ and $A(x,z)$. For 5 attributes, these distributions are tabulated in Table 9. This table is arrived at in the same manner as Tables 6 and 7, and follows their format, so no further explanation should be needed. Notice how high values of $A(x,y)$ tend to make $A(x,z)$ and $A(y,z)$ take on almost the same values. This is an immediate consequence of the fact that d_R is a metric.

	0.00	1.00	2.00	3.00	4.00	5.00		
0	0.40	0.40	0.16	0.03	0.00	0.00	0.83	0.53
1	0.24	0.43	0.25	0.07	0.01	0.00	1.17	0.63
2	0.14	0.38	0.34	0.12	0.02	0.00	1.50	0.71
3	0.09	0.30	0.37	0.20	0.05	0.00	1.83	0.78
4	0.05	0.22	0.35	0.27	0.09	0.01	2.17	0.84
5	0.03	0.16	0.31	0.31	0.16	0.03	2.50	0.88

Table 8. Conditional distribution for A(x,z) for indicated values of A(x,y) for n = 5 attributes. See text for explanation.

		0.00	1.00	2.00	3.00	4.00	5.00		
0	5	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.24
1	5	0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.63
2	5	0.00	0.00	1.00	0.00	0.00	0.00	2.00	0.90
3	5	0.00	0.00	0.00	1.00	0.00	0.00	3.00	0.98
4	5	0.00	0.00	0.00	0.00	1.00	0.00	4.00	1.00
0	4	0.80	0.20	0.00	0.00	0.00	0.00	0.20	0.32
1	4	0.05	0.76	0.19	0.00	0.00	0.00	1.14	0.66
2	4	0.00	0.12	0.71	0.18	0.00	0.00	2.06	0.88
3	4	0.00	0.00	0.23	0.62	0.15	0.00	2.92	0.97
0	3	0.64	0.32	0.04	0.00	0.00	0.00	0.40	0.39
1	3	0.09	0.59	0.28	0.04	0.00	0.00	1.26	0.68
2	3	0.01	0.23	0.51	0.23	0.03	0.00	2.04	0.85
0	2	0.51	0.38	0.10	0.01	0.00	0.00	0.60	0.46
1	2	0.15	0.47	0.30	0.07	0.01	0.00	1.32	0.68
0	1	0.41	0.41	0.15	0.03	0.00	0.00	0.80	0.52

Table 9. Conditional distributions for A(y,z) for indicated values of A(x,y) and A(x,z) on 5 attributes. See text for explanation.

§6. Distribution of Jaccard's coefficient. This coefficient is harder to work with than the others. Sokal and Sneath [11,p.131] state that the Jaccard coefficient does not define a metric. Anderberg ([1], p.117) writes that Majone and Sanday [10] "claim that the complement of the Jaccard coefficient ... is also a metric." This seems to imply that there might be some doubt on this issue. In view of the fact that Majone and Sanday's paper is not that readily available, it seems appropriate to supply a proof.

*Theorem 4. *The complement* d_J *of Jaccard's coefficient is a metric.*

Proof: It clearly suffices to establish that for arbitrary x,y,z, it is true that

$$|d_J(y,z) - d_J(x,z)| \le d_J(x,y) .$$

There is no loss in generality in assuming that $d_J(y,z) \ge d_J(x,z)$. As in §1, we agree to let a denote the number of attributes on which x and y have value 1, c the number of common 0's, and b the number of mismatches. Let $a1,b1,c1$ denote the corresponding quantities for $\{x,z\}$ and $a2,b2,c2$ those for $\{y,z\}$. Thus

(1) $d_J(x,y) = b/(a+b)$, $d_J(x,z) = b1/(a1+b1)$, $d_J(y,z) = b2/(a2+b2)$.

We are to establish that

(2) $$\frac{b2}{a2+b2} - \frac{b1}{a1+b1} \le \frac{b}{a+b} .$$

We shall proceed by induction on b, noting first that the possible attributes are as in Table 10.

*See note at end of paper.

Table 10. Possible attributes on $\{x,y,z\}$.

No. Attributes	x	y	z	xy	yz	xz
j1	1	1	1	a	a2	a1
j2	1	1	0	a	b2	b1
j3	1	0	1	b	b2	a1
j4	1	0	0	b	c2	b1
j5	0	1	1	b	a2	b1
j6	0	1	0	b	b2	c1
j7	0	0	1	c	b2	b1
j8	0	0	0	c	c2	c1

To begin the induction, suppose that $b = 0$, and note that $j2 = j4 = j5 = j6 = 0$. On the remaining attributes yz and xz are identical, so $d_J(y,z) = d_J(x,z)$ and (2) is trivial. Suppose then that (1) holds for all triples $\{x,y,z\}$ for which $b = k$, and assume that $b = k + 1$. The proof will be broken up into cases.

Case 1. $j5 \neq 0$. We may then replace a j5 attribute by a j2 attribute. This increases a and b2 by 1, while decreasing b and a2 by a like amount. By induction, this produces the inequality

(3) $$\frac{b2+1}{a2+b2} - \frac{b1}{a1+b1} \le \frac{b-1}{a+b}.$$

Since $b2/(a2+b2) \le (b2+1)/(a2+b2)$, and $(b-1)/(a+b) \le b/(a+b)$, (1) follows from this.

Case 2. $j4 \neq 0$. Now a j4 attribute may be replaced by a j2 attribute. This increases b2 and a by 1, while decreasing b by 1, thus producing

(4) $$\frac{b2+1}{a2+b2+1} - \frac{b1}{a1+b1} \le \frac{b-1}{a+b}.$$

Using the fact that $\frac{b2}{a2+b2} \le \frac{b2+1}{a2+b2+1}$, (1) again follows.

Case 3. $j3 \neq 0$ and $j4 = j5 = 0$. Here the trick is to replace a j3 attribute with a j1 attribute, noting that this establishes

(5) $$\frac{b2-1}{a2+b2} - \frac{b1}{a1+b1} \le \frac{b-1}{a+b}.$$

Noting that $j4 = j5 = 0$ implies that $a + b \le a2 + b2$, we may now write

$$\begin{aligned}\frac{b2}{a2+b2} - \frac{b1}{a1+b1} &= \frac{b2-1}{a2+b2} - \frac{b1}{a1+b1} + \frac{1}{a2+b2}\\ &\le \frac{b-1}{a+b} + \frac{1}{a2+b2}\\ &\le \frac{b-1}{a+b} + \frac{1}{a+b} = \frac{b}{a+b}.\end{aligned}$$

and this establishes (1).

Case 4. $j6 \neq 0$, but $j3 = j4 = j5 = 0$. In this case, $a1 = a2$, $a + b \le a2 + b2$, and $b = b2 - b1$. Hence

$$\begin{aligned}\frac{b2}{a2+b2} - \frac{b1}{a1+b1} &= \frac{b2}{a2+b2} - \frac{b1}{a2+b1} = \frac{a2(b2-b1)}{(a2+b2)(a2+b1)}\\ &= \frac{(a2)b}{(a2+b2)(a2+b1)} \le \frac{b}{a2+b2} \le \frac{b}{a+b}.\end{aligned}$$

The distribution of the Jaccard coefficient is based upon the fact that on random attribute data, the values of a,b,c follow a multinomial distribution with probabilities .25, .50 and .25, respectively. Thus for n binary attributes on $\{x,y\}$, the probability that $a = i_1$, $b = i_2$, and $c = i_3$ is given by $(n;i_1,i_2,i_3)(.25)^{i_1}(.50)^{i_2}(.25)^{i_3}$. By considering all possible values of i_1,i_2,i_3 and by combining those

cases which produce identical values of the Jaccard coefficient, it is easy to empirically calculate its distribution for any fixed value of n. This was done for $n = 10$, 25 and 40, and the results further grouped into the 11 categories shown in Table 11.

Table 11. Distribution of the Jaccard coefficient

Interval		Probabilities for indicated number of attributes		
Left end point	Right end point	10	25	40
0.0	0.1	.056	.009	.002
0.1	0.2	.161	.091	.051
0.2	0.3	.234	.280	.294
0.3	0.4	.192	.337	.423
0.4	0.5	.135	.198	.193
0.5	0.6	.144	.074	.035
0.6	0.7	.051	.010	.002
0.7	0.8	.018	.001	.000
0.8	0.9	.007	.000	.000
0.9	1.0	.000	.000	.000
1.0	1.0	.001	.000	.000

The expected value for Jaccard's coefficient on random data is easy to calculate. If one adopts the convention that if $a = b = 0$, then $a/(a + b) = 0$, one gets that the expected value is given by

$$\sum_{a+b+c=n} (a/(n-c))(n;a,b,c)(.25)^a(.50)^b(.25)^c =$$

$$\sum_{0 \le c < n} (1/(n-c))(n;n-c)(.75)^{n-c}(.25)^c \sum_{0 \le a \le n-c} a((n-c);a)(\tfrac{1}{3})^a(\tfrac{2}{3})^{n-c-a}$$

$$= \sum_{0 \le c < n} (1/(n-c))(n;n-c)(.75)^{n-c}(.25)^c(1/3)(n-c)$$

$$= \sum_{0 \le c < n} (1/3)(n;n-c)(.75)^{n-c}(.25)^c = (1/3)[1 - (.25)^n] .$$

When the number n of attributes is large, the distribution of d_J may be roughly approximated by that of X, where $\frac{X - 1/3}{\sqrt{8/27n}}$ is a standard normal variable. To see this, note that for fixed $n-c$, a follows the binomial distribution with $p = 1/3$. Hence the distribution of $a/(n-c)$ may be approximated by X, where $\frac{X - 1/3}{\sqrt{(1/3)(2/3)/(n-c)}}$ is standard normal. Now $n-c$ is binomial on n objects with $p = .75$. Hence its expected value is $.75n$, with a variance of $(3/16)n$. So going 2 standard deviations each side of $.75n$, we find that $n-c$ varies between $(\frac{3}{4} - \frac{1}{2}\sqrt{\frac{3}{n}})$ and $(\frac{3}{4} + \frac{1}{2}\sqrt{\frac{3}{n}}$. For n sufficiently large, we may assume that $n-c$ is close to $.75n$. To see how well this works, consider Table 12, where the actual distributions are compared with these approximations for $n = 40$, 50, 60 and 70. The variances also agree rather closely. Indeed, for $n = 50$, the actual variance is .005967, while the normal estimate has a variance of .005926. For $n = 60$, the figures are .004966 and .004938; for $n = 70$, they are .004253 and .004233.

	n = 40		n - 50		n = 60		n = 70	
$d_J \le .2$	.063	.061	.042	.042	.029	.028	.020	.019
$d_J \le .3$	.368	.349	.333	.348	.318	.328	.304	.314
$d_J \le .4$	.792	.781	.807	.816	.807	.836	.847	.853
$d_J \le .5$	.976	.974	.985	.985	.985	.991	.995	.995
$d_J \le .6$	.999	.999	.999	1	1	1	1	1
$d_J \le .7$	1	1	1	1	1	1	1	1

Table 12. For each value of n, the right column represents the actual probability, and the right column its normal approximation. See text for explanation.

The value of the Jaccard coefficient on a pair {x,z} is dependent on that of {x,y}. By considering all possible combinations of binary attributes, one can calculate the conditional distributions on {x,z} given those on {x,y}. To illustrate the situation, these distributions are presented in Table 13 for n = 5 attributes. It should be noted that the expected values for the Jaccard coefficient on {x,z} are not all that closely related to the value on {x,y}.

§7. The coefficient of special similarity. This coefficient was discussed by Farris ([2],[3]) and compared with the coefficient of special similarity. Based upon the type of argument that we called the "Third Question" (See §3), he concluded that the coefficient of special similarity was superior to the simple matching coefficient. I showed in some detail why ([7]) this type of reasoning is not valid, but in

a context rather different from the present one. Viewing the situation in the context of the present paper, it should be apparent that if one is to apply the reasoning based on the "Third Question", one must first necessarily rule out the possibility that the observed values of the correlation coefficients might be due to random errors in the data. For that reason, it is appropriate to examine the effect of the coefficient of special similarity on random binary data.

	0.00	0.20	0.25	0.33	0.40	0.50	0.60	0.67	0.75	0.80	1.00		
0.00	0.40	0.05	0.14	0.14	0.03	0.13	0.01	0.05	0.02	0.00	0.03	0.25	0.49
0.20	0.16	0.09	0.16	0.09	0.11	0.16	0.06	0.06	0.05	0.02	0.03	0.38	0.61
0.25	0.21	0.08	0.17	0.13	0.07	0.17	0.03	0.07	0.04	0.00	0.03	0.34	0.58
0.33	0.28	0.06	0.17	0.16	0.04	0.16	0.01	0.07	0.02	0.00	0.03	0.30	0.54
0.40	0.11	0.11	0.14	0.06	0.15	0.16	0.10	0.05	0.07	0.03	0.03	0.41	0.65
0.50	0.20	0.08	0.17	0.12	0.08	0.17	0.04	0.06	0.05	0.01	0.03	0.35	0.59
0.60	0.07	0.12	0.11	0.02	0.20	0.14	0.15	0.02	0.08	0.05	0.03	0.45	0.68
0.67	0.19	0.08	0.19	0.14	0.06	0.17	0.02	0.09	0.03	0.00	0.03	0.35	0.58
0.75	0.09	0.11	0.16	0.05	0.14	0.19	0.08	0.05	0.09	0.02	0.03	0.42	0.65
0.80	0.05	0.14	0.06	0.00	0.25	0.09	0.22	0.00	0.06	0.09	0.03	0.47	0.70
1.00	0.21	0.08	0.16	0.12	0.08	0.16	0.04	0.06	0.04	0.01	0.03	0.34	0.58

Table 13. Conditional distributions for the Jaccard coefficient on {x,z} given its value on {x,y}. See text for explanation.

Before doing this, let me examine the nature of this coefficient. For binary attribute data, one identifies one of the two states for each attribute as being "uninformative", and then uses the number of matched pairs of "informative" states as a measure of similarity. If the attribute states are recoded so as to identify each uninformative state by a 0, this then reduces to the coefficient of Russell and Rao. For that reason, there is no reason to examine the distribution of this coefficient. Now this is an extremely attractive idea that holds great promise. My quarrel is not with the coefficient, but rather with the way that Farris attempted to demonstrate its superiority. What he did in both of his papers was to take the first object of the set of objects to be classified and arbitrarily specify that each of its attribute states shall be uninformative. Now this artificially imposes structure on any data set because it forces that first object to have distance 1 from every other object. The means and standard deviations for the Kendall and Spearman coefficients on binary data for various numbers of elements and attributes appear in Table 14. They should be compared with the results that were earlier obtained for the coefficient of Russell and Rao. Tables 15 and 16 contain the cutpoints at which the null hypothesis that the observed values are due to random effects may be rejected. Thus a value of .9400 for the Kendall coefficient on a 6 element set with 25 attributes could at a significance level of 90% be due to random data, while

Table 14. Values of Kendall and Spearman coefficients for the coefficient of special similarity.

No. Elements	No. Attributes	Kendall Mean	Kendall SD	Spearman Mean	Spearman SD
4	10	.9437	.0707	.9608	.0667
4	25	.9593	.0333	.9813	.0204
4	50	.9557	.0279	.9805	.0156
5	10	.8941	.0650	.9302	.0599
5	25	.9058	.0507	.9507	.0329
5	50	.8927	.0538	.9448	.0339
6	10	.8626	.0670	.9083	.0567
6	25	.8635	.0606	.9229	.0435
6	50	.8531	.0622	.9188	.0449

Each result was based upon 200 trials using random binary attribute data.

Table 15. Cutpoints for the Kendall coefficient using the coefficient of special similarity

No. Elements	No. Attributes	.80	.85	.90	.95	.99
4	10	1	1	1	1	1
4	25	1	1	1	1	1
4	50	.9574	1	1	1	1
5	10	.9416	.9448	.9710	.9726	1
5	25	.9473	.9597	.9597	.9726	.9860
5	50	.9448	.9473	.9597	.9597	.9860
6	10	.9262	.9342	.9524	.9613	.9777
6	25	.9160	.9294	.9424	.9488	.9612
6	50	.9083	.9160	.9384	.9448	.9584

Table 16. Cutpoints for Spearman coefficient using the coefficient of special similarity

No. Elements	No. Attributes	.80	.85	.90	.95	.99
4	10	1	1	1	1	1
4	25	1	1	1	1	1
4	50	.9837	1	1	1	1
5	10	.9705	.9802	.9899	.9901	1
5	25	.9837	.9869	.9869	.9901	.9967
5	50	.9802	.9837	.9869	.9869	.9967
6	10	.9572	.9658	.9783	.9839	.9954
6	25	.9627	.9686	.9774	.9821	.9888
6	50	.9564	.9671	.9773	.9831	.9878

(Based upon 200 trials with random binary data)

value of .9450 would presumably reflect some structure in the data. Notice that for all practical purposes. the null hypothesis cannot be rejected for a 4 element set. These tables should serve to illustrate the dangers inherent in using this type of argument to demonstrate the superiority of one dissimilarity coefficient over another.

§8. Conclusion. It was argued that rho and tau may both be used to evaluate ordinal cluster techniques on the same input dissimilarity coefficient, and that when properly used, either of them may serve as an indicator of the probability that an observed output reflects some actual structure as opposed to being due to random error. On the other hand, it was observed that there is little basis for using either of these coefficients as a measure of how well an intermediate dissimilarity coefficient reflects the structure contained in binary attribute data. On random data, evidence was presented that seemed to indicate that special similarity imposed the most structure, followed in decreasing order by Russell and Rao, Jaccard, and the simple matching coefficient. The distributions and properties of these coefficients are examined, and a disadvantage of the simple matching coefficient established. More work along these lines needs to be done, but the present work serves to point out that some caution and understanding needs to be exervised before using any of these coefficients.

In closing, it seems worth mentioning that once one understands the distribution of a dissimilarity coefficient on random data, it should then be possible to construct confidence intervals for the observed values of that coefficient, and use these values to decide whether the observed values represent actual structure or might be due to some sort of noise. This will be explored in a later paper.

NOTE: The author has learned from Jean-Pierre Croteau that the result announced in Theorem 4 also appears in Pagès, J. P., Cailliez, Introduction à l'analyse des dennées.

REFERENCES

[1] M. R. Anderberg, Cluster Analysis for Applications, Academic Press (1973), 359 pp.

[2] J. S. Farris, On the phenetic approach to vertebrate classification. In M. K. Hecht, P. C. Goody and B. M. Hecht (Eds.), "Major patterns in vertebrate evolution". Plenum (1977), pp. 823-850.

[3] ________, On the naturalness of phylogenetic classifications, Syst. Zool 28 (1979), pp. 200-214.

[4] D. W. Goodall, The distribution of the matching coefficient. Biometrics 23 (1967), pp. 647-656.

[5] L. J. Hubert and F. B. Baker, An empirical comparison of baseline models for goodness-of-fit in r-diameter hierarchical clustering. In J. van Ryzin (Ed), "Classification and clustering". Academic Press (1977), pp. 131-153.

[6] M. F. Janowitz, Monotone equivariant cluster methods, SIAM J. Appl. Math. 37 (1979), pp. 148-165.

[7] ___________, Similarity measures on binary data, Syst. Zool. 29 (1980), pp. 342-359.

[8] ___________, Optimality measures for monotone equivariant cluster techniques, University of Massachusetts Technical Report J8001, 32 pp.

[9] M. G. Kendall, Rank order correlation methods, Griffin (1970) 4th Edition, 202 pp.

[10] G. Majone and P. R. Sanday, On the numerical classification of nominal data, Rep. No. RR-118. AD 665006. Graduate School of Ind. Administration, Carnegie-Mellon University, Pittsburgh, PA

[11] P. H. A. Sneath and R. P. Sokal, Numerical Taxonomy, Freeman (1973), 573 pp.

APPENDIX

HISTOGRAM FOR VARIABLE KENJ50

INTERVAL	FR	PCT	I +70 +140 +210
0.2670≤X< 0.2996	4	0.8	I□
0.2996≤X< 0.3323	0	0.0	I
0.3323≤X< 0.3649	0	0.0	I
0.3649≤X< 0.3976	61	12.2	I□□□□□□□□□
0.3976≤X< 0.4302	6	1.2	I□
0.4302≤X< 0.4629	4	0.8	I□
0.4629≤X< 0.4955	5	1.0	I□
0.4955≤X< 0.5282	0	0.0	I
0.5282≤X< 0.5608	64	12.8	I□□□□□□□□□
0.5608≤X< 0.5935	4	0.8	I□
0.5935≤X< 0.6261	19	3.8	I□□□
0.6261≤X< 0.6588	4	0.8	I□
0.6588≤X< 0.6914	0	0.0	I
0.6914≤X< 0.7241	93	18.6	I□□□□□□□□□□□□□
0.7241≤X< 0.7567	7	1.4	I□
0.7567≤X< 0.7894	62	12.4	I□□□□□□□□□
0.7894≤X< 0.8220	19	3.8	I□□□
0.8220≤X< 0.8547	3	0.6	I
0.8547≤X< 0.8873	143	28.7	I□□□□□□□□□□□□□□□□□□□□
0.8873≤X≤ 0.9200	1	0.2	I

NUMBER OF MISSING CASES: 1
ONE □ REPRESENTS 7 CASES.

THERE ARE 4 OBJECTS TO BE CLASSIFIED

HISTOGRAM FOR VARIABLE KENS50

INTERVAL	FR	PCT	I +40 +80 +120
0.2670≤X< 0.3036	16	3.2	I□□□□
0.3036≤X< 0.3403	0	0.0	I
0.3403≤X< 0.3769	3	0.6	I□
0.3769≤X< 0.4136	37	7.5	I□□□□□□□□□
0.4136≤X< 0.4502	7	1.4	I□□
0.4502≤X< 0.4869	13	2.6	I□□□
0.4869≤X< 0.5235	2	0.4	I□
0.5235≤X< 0.5602	31	6.3	I□□□□□□□□
0.5602≤X< 0.5968	14	2.8	I□□□□
0.5968≤X< 0.6335	70	14.1	I□□□□□□□□□□□□□□□□□□
0.6335≤X< 0.6701	40	8.1	I□□□□□□□□□□
0.6701≤X< 0.7068	28	5.7	I□□□□□□□
0.7068≤X< 0.7434	11	2.2	I□□□
0.7434≤X< 0.7801	30	6.1	I□□□□□□□□
0.7801≤X< 0.8167	82	16.6	I□□□□□□□□□□□□□□□□□□□□□
0.8167≤X< 0.8534	21	4.2	I□□□□□
0.8534≤X< 0.8900	78	15.8	I□□□□□□□□□□□□□□□□□□□□
0.8900≤X< 0.9267	10	2.0	I□□□
0.9267≤X< 0.9633	0	0.0	I
0.9633≤X≤ 1.0000	2	0.4	I□

NUMBER OF MISSING CASES: 5
ONE □ REPRESENTS 4 CASES.

THERE ARE 4 OBJECTS TO BE CLASSIFIED

```
HISTOGRAM FOR VARIABLE KENR50

   INTERVAL          FR   PCT   I          +40        +80        +120
0.2670≤X< 0.3036      6   1.2   I□□
0.3036≤X< 0.3403      0   0.0   I
0.3403≤X< 0.3769      5   1.0   I□
0.3769≤X< 0.4136     14   2.9   I□□□□
0.4136≤X< 0.4502     12   2.5   I□□□
0.4502≤X< 0.4869     12   2.5   I□□□
0.4869≤X< 0.5235      8   1.7   I□□
0.5235≤X< 0.5602     13   2.7   I□□□
0.5602≤X< 0.5968     12   2.5   I□□□
0.5968≤X< 0.6335     61  12.7   I□□□□□□□□□□□□□□□
0.6335≤X< 0.6701     46   9.6   I□□□□□□□□□□□□
0.6701≤X< 0.7068     36   7.5   I□□□□□□□□□
0.7068≤X< 0.7434     21   4.4   I□□□□□
0.7434≤X< 0.7801      8   1.7   I□□
0.7801≤X< 0.8167     70  14.6   I□□□□□□□□□□□□□□□□□□
0.8167≤X< 0.8534     36   7.5   I□□□□□□□□□
0.8534≤X< 0.8900     95  19.8   I□□□□□□□□□□□□□□□□□□□□□□□□
0.8900≤X< 0.9267     18   3.7   I□□□□□
0.9267≤X< 0.9633      4   0.8   I□
0.9633≤X≤ 1.0000      4   0.8   I□

NUMBER OF MISSING CASES: 19
ONE □ REPRESENTS 4 CASES.

THERE ARE 4 OBJECTS TO BE CLASSIFIED
```

```
HISTOGRAM FOR VARIABLE SPR50

   INTERVAL          FR   PCT   I          +40        +80
0.2970≤X< 0.3321      6   1.2   I□□
0.3321≤X< 0.3673      0   0.0   I
0.3673≤X< 0.4024      5   1.0   I□
0.4024≤X< 0.4376      1   0.2   I
0.4376≤X< 0.4727     13   2.7   I□□□
0.4727≤X< 0.5079     14   2.9   I□□□□
0.5079≤X< 0.5430      1   0.2   I
0.5430≤X< 0.5782     17   3.5   I□□□□
0.5782≤X< 0.6133      5   1.0   I□
0.6133≤X< 0.6485     20   4.2   I□□□□□
0.6485≤X< 0.6836     57  11.9   I□□□□□□□□□□□□□□
0.6836≤X< 0.7188     61  12.7   I□□□□□□□□□□□□□□□
0.7188≤X< 0.7539      9   1.9   I□□
0.7539≤X< 0.7891     26   5.4   I□□□□□□□
0.7891≤X< 0.8242     37   7.7   I□□□□□□□□□
0.8242≤X< 0.8594     12   2.5   I□□□
0.8594≤X< 0.8945     60  12.5   I□□□□□□□□□□□□□□□
0.8945≤X< 0.9297     53  11.0   I□□□□□□□□□□□□□
0.9297≤X< 0.9648     76  15.8   I□□□□□□□□□□□□□□□□□□□
0.9648≤X≤ 1.0000      8   1.7   I□□

NUMBER OF MISSING CASES: 19
ONE □ REPRESENTS 4 CASES.

THERE ARE 4 OBJECTS TO BE CLASSIFIED
```



```
HISTOGRAM FOR VARIABLE SPJ50

   INTERVAL          FR   PCT   I          +60        +120       +180
0.2970≤X< 0.329       4   0.8   I□
0.329≤X< 0.362        0   0.0   I
0.362≤X< 0.395        0   0.0   I
0.395≤X< 0.428        0   0.0   I
0.428≤X< 0.461        0   0.0   I
0.461≤X< 0.493       67  13.4   I□□□□□□□□□□□
0.493≤X< 0.526        4   0.8   I□
0.526≤X< 0.559        5   1.0   I□
0.559≤X< 0.592        0   0.0   I
0.592≤X< 0.625       64  12.8   I□□□□□□□□□□□
0.625≤X< 0.657        5   1.0   I□
0.657≤X< 0.690       15   3.0   I□□□
0.690≤X< 0.723        7   1.4   I□
0.723≤X< 0.756        1   0.2   I
0.756≤X< 0.789       95  19.0   I□□□□□□□□□□□□□□□□
0.789≤X< 0.821        3   0.6   I□
0.821≤X< 0.854       63  12.6   I□□□□□□□□□□□
0.854≤X< 0.887       19   3.8   I□□□
0.887≤X< 0.920        4   0.8   I□
0.920≤X≤ 0.953      143  28.7   I□□□□□□□□□□□□□□□□□□□□□□□□

NUMBER OF MISSING CASES: 1
ONE □ REPRESENTS 6 CASES.
```

```
HISTOGRAM FOR VARIABLE SPS50

   INTERVAL          FR   PCT   I          +30        +60        +90
0.2970≤X< 0.3321     16   3.2   I□□□□□
0.3321≤X< 0.3673      0   0.0   I
0.3673≤X< 0.4024      3   0.6   I□
0.4024≤X< 0.4376      2   0.4   I□
0.4376≤X< 0.4727     35   7.1   I□□□□□□□□□□□□
0.4727≤X< 0.5079      9   1.8   I□□□
0.5079≤X< 0.5430      0   0.0   I
0.5430≤X< 0.5782     13   2.6   I□□□□
0.5782≤X< 0.6133      6   1.2   I□□
0.6133≤X< 0.6485     39   7.9   I□□□□□□□□□□□□□
0.6485≤X< 0.6836     66  13.3   I□□□□□□□□□□□□□□□□□□□□□□
0.6836≤X< 0.7188     46   9.3   I□□□□□□□□□□□□□□□
0.7188≤X< 0.7539      2   0.4   I□
0.7539≤X< 0.7891     36   7.3   I□□□□□□□□□□□□
0.7891≤X< 0.8242      5   1.0   I□□
0.8242≤X< 0.8594     65  13.1   I□□□□□□□□□□□□□□□□□□□□□□
0.8594≤X< 0.8945     58  11.7   I□□□□□□□□□□□□□□□□□□□
0.8945≤X< 0.9297     54  10.9   I□□□□□□□□□□□□□□□□□□
0.9297≤X< 0.9648     38   7.7   I□□□□□□□□□□□□□
0.9648≤X≤ 1.0000      2   0.4   I□

NUMBER OF MISSING CASES: 5
ONE □ REPRESENTS 3 CASES.

THERE ARE 4 OBJECTS TO BE CLASSIFIED
```

```
HISTOGRAM FOR VARIABLE KENR50

      INTERVAL              FR   PCT   I              +40          +80
0.15100≤X< 0.19005          2   0.4   I□
0.19005≤X< 0.22910          2   0.4   I□
0.22910≤X< 0.26815          2   0.4   I□
0.26815≤X< 0.30720          5   1.0   I□
0.30720≤X< 0.34625          4   0.8   I□
0.34625≤X< 0.38530          8   1.6   I□□
0.38530≤X< 0.42435         12   2.4   I□□□
0.42435≤X< 0.46340         18   3.6   I□□□□□
0.46340≤X< 0.50245         42   8.5   I□□□□□□□□□□□
0.50245≤X< 0.54150         34   6.9   I□□□□□□□□□
0.54150≤X< 0.58055         29   5.9   I□□□□□□□
0.58055≤X< 0.61960         20   4.0   I□□□□□
0.61960≤X< 0.65865         42   8.5   I□□□□□□□□□□□
0.65865≤X< 0.69770         54  10.9   I□□□□□□□□□□□□□□
0.69770≤X< 0.73675         43   8.7   I□□□□□□□□□□□
0.73675≤X< 0.77580         68  13.7   I□□□□□□□□□□□□□□□□□
0.77580≤X< 0.81485         37   7.5   I□□□□□□□□□
0.81485≤X< 0.85390         36   7.3   I□□□□□□□□□
0.85390≤X< 0.89295         29   5.9   I□□□□□□□
0.89295≤X≤ 0.93200          8   1.6   I□□

NUMBER OF MISSING CASES: 5
ONE □ REPRESENTS 4 CASES.

THERE ARE 5 OBJECTS TO BE CLASSIFIED
```

```
HISTOGRAM FOR VARIABLE KENS50

      INTERVAL          FR   PCT   I              +30          +60          +90
0.1320≤X< 0.1710         1   0.2   I
0.1710≤X< 0.2101         3   0.6   I□
0.2101≤X< 0.2491         4   0.8   I□
0.2491≤X< 0.2882         4   0.8   I□
0.2882≤X< 0.3272         9   1.8   I□□□
0.3272≤X< 0.3663        15   3.0   I□□□□□
0.3663≤X< 0.4053        15   3.0   I□□□□□
0.4053≤X< 0.4444        19   3.8   I□□□□□□
0.4444≤X< 0.4834        42   8.4   I□□□□□□□□□□□□□□
0.4834≤X< 0.5225        40   8.0   I□□□□□□□□□□□□□
0.5225≤X< 0.5615        33   6.6   I□□□□□□□□□□□
0.5615≤X< 0.6006        35   7.0   I□□□□□□□□□□□□
0.6006≤X< 0.6396        66  13.2   I□□□□□□□□□□□□□□□□□□□□□□
0.6396≤X< 0.6787        55  11.0   I□□□□□□□□□□□□□□□□□□
0.6787≤X< 0.7177        38   7.6   I□□□□□□□□□□□□□
0.7177≤X< 0.7568        36   7.2   I□□□□□□□□□□□□
0.7568≤X< 0.7958        28   5.6   I□□□□□□□□□
0.7958≤X< 0.8349        41   8.2   I□□□□□□□□□□□□□□
0.8349≤X< 0.8739         8   1.6   I□□□
0.8739≤X≤ 0.9130         8   1.6   I□□□

NUMBER OF MISSING CASES: 0
ONE □ REPRESENTS 3 CASES.

THERE ARE 5 OBJECTS TO BE CLASSIFIED
```



```
HISTOGRAM FOR VARIABLE KENJ50

      INTERVAL          FR   PCT   I              +30          +60          +90
0.1260≤X< 0.1638         3   0.6   I□
0.1638≤X< 0.2016         0   0.0   I
0.2016≤X< 0.2394         7   1.4   I□□
0.2394≤X< 0.2772        10   2.0   I□□□
0.2772≤X< 0.3150         5   1.0   I□□
0.3150≤X< 0.3528        12   2.4   I□□□□
0.3528≤X< 0.3906        11   2.2   I□□□□
0.3906≤X< 0.4284        23   4.6   I□□□□□□□□
0.4284≤X< 0.4662         9   1.8   I□□□
0.4662≤X< 0.5040        36   7.2   I□□□□□□□□□□□□
0.5040≤X< 0.5418        44   8.8   I□□□□□□□□□□□□□□□
0.5418≤X< 0.5796         9   1.8   I□□□
0.5796≤X< 0.6174        51  10.2   I□□□□□□□□□□□□□□□□□
0.6174≤X< 0.6552        69  13.8   I□□□□□□□□□□□□□□□□□□□□□□□
0.6552≤X< 0.6930        45   9.0   I□□□□□□□□□□□□□□□
0.6930≤X< 0.7308        21   4.2   I□□□□□□□
0.7308≤X< 0.7686        56  11.2   I□□□□□□□□□□□□□□□□□□□
0.7686≤X< 0.8064        47   9.4   I□□□□□□□□□□□□□□□□
0.8064≤X< 0.8442        27   5.4   I□□□□□□□□□
0.8442≤X≤ 0.8820        15   3.0   I□□□□□

NUMBER OF MISSING CASES: 0
ONE □ REPRESENTS 3 CASES.

THERE ARE 5 OBJECTS TO BE CLASSIFIED
```

```
HISTOGRAM FOR VARIABLE KPERM

      INTERVAL          FR   PCT   I              +30          +60          +90
0.1760≤X< 0.2113         4   0.8   I□
0.2113≤X< 0.2466        12   2.4   I□□□□
0.2466≤X< 0.2819        19   3.8   I□□□□□□
0.2819≤X< 0.3172         2   0.4   I□
0.3172≤X< 0.3525        25   5.0   I□□□□□□□□
0.3525≤X< 0.3878        35   7.0   I□□□□□□□□□□□□
0.3878≤X< 0.4231        14   2.8   I□□□□□
0.4231≤X< 0.4584        34   6.8   I□□□□□□□□□□□
0.4584≤X< 0.4937        52  10.4   I□□□□□□□□□□□□□□□□□
0.4937≤X< 0.5290        17   3.4   I□□□□□□
0.5290≤X< 0.5643        44   8.8   I□□□□□□□□□□□□□□□
0.5643≤X< 0.5996        70  14.0   I□□□□□□□□□□□□□□□□□□□□□□□
0.5996≤X< 0.6349        26   5.2   I□□□□□□□□□
0.6349≤X< 0.6702        31   6.2   I□□□□□□□□□□
0.6702≤X< 0.7055        55  11.0   I□□□□□□□□□□□□□□□□□□
0.7055≤X< 0.7408        10   2.0   I□□□
0.7408≤X< 0.7761        16   3.2   I□□□□□
0.7761≤X< 0.8114        24   4.8   I□□□□□□□□
0.8114≤X< 0.8467         5   1.0   I□□
0.8467≤X≤ 0.8820         5   1.0   I□□

NUMBER OF MISSING CASES: 0
ONE □ REPRESENTS 3 CASES.

THERE ARE 5 OBJECTS TO BE CLASSIFIED
```

```
HISTOGRAM FOR VARIABLE SPS50

    INTERVAL          FR   PCT   I           +30        +60        +90
0.1460≤X< 0.1869      1   0.2   I
0.1869≤X< 0.2279      0   0.0   I
0.2279≤X< 0.2688      3   0.6   I□
0.2688≤X< 0.3098      6   1.2   I□□
0.3098≤X< 0.3507      6   1.2   I□□
0.3507≤X< 0.3917      8   1.6   I□□□
0.3917≤X< 0.4326     17   3.4   I□□□□□□
0.4326≤X< 0.4736     14   2.8   I□□□□□
0.4736≤X< 0.5145     28   5.6   I□□□□□□□□□
0.5145≤X< 0.5555     42   8.4   I□□□□□□□□□□□□□□
0.5555≤X< 0.5964     34   6.8   I□□□□□□□□□□□
0.5964≤X< 0.6374     28   5.6   I□□□□□□□□□
0.6374≤X< 0.6783     48   9.6   I□□□□□□□□□□□□□□□□
0.6783≤X< 0.7193     63  12.6   I□□□□□□□□□□□□□□□□□□□□□
0.7193≤X< 0.7602     49   9.8   I□□□□□□□□□□□□□□□□
0.7602≤X< 0.8012     39   7.8   I□□□□□□□□□□□□□
0.8012≤X< 0.8421     36   7.2   I□□□□□□□□□□□□
0.8421≤X< 0.8831     38   7.6   I□□□□□□□□□□□□□
0.8831≤X< 0.9240     28   5.6   I□□□□□□□□□
0.9240≤X≤ 0.9650     12   2.4   I□□□□

NUMBER OF MISSING CASES: 0
ONE □ REPRESENTS 3 CASES.

THERE ARE 5 OBJECTS TO BE CLASSIFIED
```

```
HISTOGRAM FOR VARIABLE SPR50

    INTERVAL          FR   PCT   I           +30        +60
0.1870≤X< 0.2262      3   0.6   I□
0.2262≤X< 0.2654      1   0.2   I
0.2654≤X< 0.3046      3   0.6   I□
0.3046≤X< 0.3438      3   0.6   I□
0.3438≤X< 0.3830      5   1.0   I□□
0.3830≤X< 0.4222      4   0.8   I□
0.4222≤X< 0.4614     11   2.2   I□□□□
0.4614≤X< 0.5006     11   2.2   I□□□□
0.5006≤X< 0.5398     30   6.1   I□□□□□□□□□□
0.5398≤X< 0.5790     38   7.7   I□□□□□□□□□□□□□
0.5790≤X< 0.6182     27   5.5   I□□□□□□□□□
0.6182≤X< 0.6574     23   4.6   I□□□□□□□□
0.6574≤X< 0.6966     22   4.4   I□□□□□□□
0.6966≤X< 0.7358     50  10.1   I□□□□□□□□□□□□□□□□□
0.7358≤X< 0.7750     40   8.1   I□□□□□□□□□□□□□
0.7750≤X< 0.8142     48   9.7   I□□□□□□□□□□□□□□□□
0.8142≤X< 0.8534     59  11.9   I□□□□□□□□□□□□□□□□□□□□
0.8534≤X< 0.8926     55  11.1   I□□□□□□□□□□□□□□□□□□
0.8926≤X< 0.9318     33   6.7   I□□□□□□□□□□□
0.9318≤X≤ 0.9710     29   5.9   I□□□□□□□□□□

NUMBER OF MISSING CASES: 5
ONE □ REPRESENTS 3 CASES.

THERE ARE 5 OBJECTS TO BE CLASSIFIED
```



```
HISTOGRAM FOR VARIABLE SPJ50

    INTERVAL          FR   PCT   I           +30        +60
0.1970≤X< 0.2348      3   0.6   I□
0.2348≤X< 0.2726      2   0.4   I□
0.2726≤X< 0.3104      5   1.0   I□□
0.3104≤X< 0.3482      7   1.4   I□□
0.3482≤X< 0.3860     14   2.8   I□□□□□
0.3860≤X< 0.4238     10   2.0   I□□□
0.4238≤X< 0.4616      5   1.0   I□□
0.4616≤X< 0.4994     16   3.2   I□□□□□
0.4994≤X< 0.5372     20   4.0   I□□□□□□□
0.5372≤X< 0.5750     24   4.8   I□□□□□□□□
0.5750≤X< 0.6128     35   7.0   I□□□□□□□□□□□□
0.6128≤X< 0.6506     27   5.4   I□□□□□□□□□
0.6506≤X< 0.6884     38   7.6   I□□□□□□□□□□□□□
0.6884≤X< 0.7262     43   8.6   I□□□□□□□□□□□□□□
0.7262≤X< 0.7640     52  10.4   I□□□□□□□□□□□□□□□□□
0.7640≤X< 0.8018     48   9.6   I□□□□□□□□□□□□□□□□
0.8018≤X< 0.8396     39   7.8   I□□□□□□□□□□□□□
0.8396≤X< 0.8774     44   8.8   I□□□□□□□□□□□□□□□
0.8774≤X< 0.9152     49   9.8   I□□□□□□□□□□□□□□□□
0.9152≤X≤ 0.9530     19   3.8   I□□□□□□

NUMBER OF MISSING CASES: 0
ONE □ REPRESENTS 3 CASES.

THERE RAE 5 OBJECTS TO BE CLASSIFIED
```

```
HISTOGRAM FOR VARIABLE SPERM

    INTERVAL          FR   PCT   I           +30        +60        +90
0.2290≤X< 0.2652      4   0.8   I□
0.2652≤X< 0.3014     10   2.0   I□□□
0.3014≤X< 0.3376     13   2.6   I□□□□
0.3376≤X< 0.3738     11   2.2   I□□□□
0.3738≤X< 0.4100     24   4.8   I□□□□□□□□
0.4100≤X< 0.4462     18   3.6   I□□□□□□
0.4462≤X< 0.4824     34   6.8   I□□□□□□□□□□□
0.4824≤X< 0.5186     20   4.0   I□□□□□□□
0.5186≤X< 0.5548     29   5.8   I□□□□□□□□□□
0.5548≤X< 0.5910     44   8.8   I□□□□□□□□□□□□□□□
0.5910≤X< 0.6272     44   8.8   I□□□□□□□□□□□□□□□
0.6272≤X< 0.6634     15   3.0   I□□□□□
0.6634≤X< 0.6996     61  12.2   I□□□□□□□□□□□□□□□□□□□□
0.6996≤X< 0.7358     46   9.2   I□□□□□□□□□□□□□□□
0.7358≤X< 0.7720     18   3.6   I□□□□□□
0.7720≤X< 0.8082     49   9.8   I□□□□□□□□□□□□□□□□
0.8082≤X< 0.8444     19   3.8   I□□□□□□
0.8444≤X< 0.8806     14   2.8   I□□□□□
0.8806≤X< 0.9168     19   3.8   I□□□□□□
0.9168≤X≤ 0.9530      8   1.6   I□□□

NUMBER OF MISSING CASES: 0
ONE □ REPRESENTS 3 CASES.

THERE ARE 5 OBJECTS TO BE CLASSIFIED
```

HISTOGRAM FOR VARIABLE KENS50

INTERVAL	FR	PCT	I ↓30 ↓60
0.0750≤X< 0.1143	4	0.8	I□
0.1143≤X< 0.1537	2	0.4	I□
0.1537≤X< 0.1930	6	1.2	I□□
0.1930≤X< 0.2324	6	1.2	I□□
0.2324≤X< 0.2717	12	2.4	I□□□□
0.2717≤X< 0.3111	19	3.8	I□□□□□□
0.3111≤X< 0.3504	16	3.2	I□□□□□
0.3504≤X< 0.3898	21	4.2	I□□□□□□□
0.3898≤X< 0.4291	38	7.6	I□□□□□□□□□□□□□
0.4291≤X< 0.4685	45	9.0	I□□□□□□□□□□□□□□□
0.4685≤X< 0.5078	47	9.4	I□□□□□□□□□□□□□□□□
0.5078≤X< 0.5472	58	11.6	I□□□□□□□□□□□□□□□□□□□
0.5472≤X< 0.5865	39	7.8	I□□□□□□□□□□□□□
0.5865≤X< 0.6259	54	10.8	I□□□□□□□□□□□□□□□□□□
0.6259≤X< 0.6652	45	9.0	I□□□□□□□□□□□□□□□
0.6652≤X< 0.7046	34	6.8	I□□□□□□□□□□□
0.7046≤X< 0.7439	30	6.0	I□□□□□□□□□□
0.7439≤X< 0.7833	12	2.4	I□□□□
0.7833≤X< 0.8226	6	1.2	I□□
0.8226≤X≤ 0.8620	5	1.0	I□□

NUMBER OF MISSING CASES: 1
ONE □ REPRESENTS 3 CASES.

THERE ARE 6 OBJECTS TO BE CLASSIFIED

HISTOGRAM FOR VARIABLE KENR50

INTERVAL	FR	PCT	I ↓30 ↓60
0.1550≤X< 0.1937	3	0.6	I□
0.1937≤X< 0.2325	1	0.2	I
0.2325≤X< 0.2712	1	0.2	I
0.2712≤X< 0.3100	6	1.2	I□□
0.3100≤X< 0.3487	9	1.8	I□□□
0.3487≤X< 0.3875	13	2.6	I□□□□
0.3875≤X< 0.4262	18	3.6	I□□□□□□
0.4262≤X< 0.4650	24	4.8	I□□□□□□□□
0.4650≤X< 0.5037	30	6.0	I□□□□□□□□□□
0.5037≤X< 0.5425	35	7.0	I□□□□□□□□□□□□
0.5425≤X< 0.5812	43	8.6	I□□□□□□□□□□□□□□
0.5812≤X< 0.6200	43	8.6	I□□□□□□□□□□□□□□
0.6200≤X< 0.6587	55	11.0	I□□□□□□□□□□□□□□□□□□
0.6587≤X< 0.6975	57	11.4	I□□□□□□□□□□□□□□□□□□□
0.6975≤X< 0.7362	47	9.4	I□□□□□□□□□□□□□□□□
0.7362≤X< 0.7750	43	8.6	I□□□□□□□□□□□□□□
0.7750≤X< 0.8137	30	6.0	I□□□□□□□□□□
0.8137≤X< 0.8525	25	5.0	I□□□□□□□□
0.8525≤X< 0.8912	13	2.6	I□□□□
0.8912≤X≤ 0.9300	3	0.6	I□

NUMBER OF MISSING CASES: 1
ONE □ REPRESENTS 3 CASES.

THERE ARE 6 OBJECTS TO BE CLASSIFIED

HISTOGRAM FOR VARIABLE KENJ50

INTERVAL	FR	PCT	I ↓30 ↓60
0.1040≤X< 0.1427	3	0.6	I□
0.1427≤X< 0.1815	0	0.0	I
0.1815≤X< 0.2202	4	0.8	I□
0.2202≤X< 0.2590	8	1.6	I□□□
0.2590≤X< 0.2977	7	1.4	I□□
0.2977≤X< 0.3365	17	3.4	I□□□□□□
0.3365≤X< 0.3752	20	4.0	I□□□□□□□
0.3752≤X< 0.4140	37	7.4	I□□□□□□□□□□□□
0.4140≤X< 0.4527	37	7.4	I□□□□□□□□□□□□
0.4527≤X< 0.4915	46	9.2	I□□□□□□□□□□□□□□□
0.4915≤X< 0.5302	38	7.6	I□□□□□□□□□□□□□
0.5302≤X< 0.5690	51	10.2	I□□□□□□□□□□□□□□□□□
0.5690≤X< 0.6077	46	9.2	I□□□□□□□□□□□□□□□
0.6077≤X< 0.6465	53	10.6	I□□□□□□□□□□□□□□□□□□
0.6465≤X< 0.6852	40	8.0	I□□□□□□□□□□□□□
0.6852≤X< 0.7240	33	6.6	I□□□□□□□□□□□
0.7240≤X< 0.7627	25	5.0	I□□□□□□□□
0.7627≤X< 0.8015	18	3.6	I□□□□□□
0.8015≤X< 0.8402	13	2.6	I□□□□
0.8402≤X≤ 0.8790	4	0.8	I□

NUMBER OF MISSING CASES: 0
ONE □ REPRESENTS 3 CASES.

THERTE ARE 6 OBJECTS TO BE CLASSIFIED

HISTOGRAM FOR VARIABLE KFERN

INTERVAL	FR	PCT	I ↓30 ↓60
0.1160≤X< 0.1496	4	0.8	I□
0.1496≤X< 0.1832	8	1.6	I□□□
0.1832≤X< 0.2168	9	1.8	I□□□
0.2168≤X< 0.2504	20	4.0	I□□□□□□□
0.2504≤X< 0.2840	24	4.8	I□□□□□□□□
0.2840≤X< 0.3176	29	5.8	I□□□□□□□□□□
0.3176≤X< 0.3512	31	6.2	I□□□□□□□□□□
0.3512≤X< 0.3848	33	6.6	I□□□□□□□□□□□
0.3848≤X< 0.4184	48	9.6	I□□□□□□□□□□□□□□□□
0.4184≤X< 0.4520	52	10.4	I□□□□□□□□□□□□□□□□□
0.4520≤X< 0.4856	46	9.2	I□□□□□□□□□□□□□□□
0.4856≤X< 0.5192	41	8.2	I□□□□□□□□□□□□□□
0.5192≤X< 0.5528	26	5.2	I□□□□□□□□□
0.5528≤X< 0.5864	40	8.0	I□□□□□□□□□□□□□
0.5864≤X< 0.6200	21	4.2	I□□□□□□□
0.6200≤X< 0.6536	21	4.2	I□□□□□□□
0.6536≤X< 0.6872	19	3.8	I□□□□□□
0.6872≤X< 0.7208	14	2.8	I□□□□□
0.7208≤X< 0.7544	10	2.0	I□□□
0.7544≤X≤ 0.7880	4	0.8	I□

NUMBER OF MISSING CASES: 0
ONE □ REPRESENTS 3 CASES.

THERE ARE 6 OBJECTS TO BE CLASSIFIED

```
HISTOGRAM FOR VARIABLE SPS50

   INTERVAL          FR  PCT   I          +30         +60         +90
0.0980≤X< 0.1398      5  1.0   I□□
0.1398≤X< 0.1816      2  0.4   I□
0.1816≤X< 0.2234      4  0.8   I□
0.2234≤X< 0.2652      7  1.4   I□□
0.2652≤X< 0.3070      9  1.8   I□□□
0.3070≤X< 0.3488     15  3.0   I□□□□□
0.3488≤X< 0.3906     15  3.0   I□□□□□
0.3906≤X< 0.4324     16  3.2   I□□□□□
0.4324≤X< 0.4742     30  6.0   I□□□□□□□□□□
0.4742≤X< 0.5160     42  8.4   I□□□□□□□□□□□□□□
0.5160≤X< 0.5578     41  8.2   I□□□□□□□□□□□□□□
0.5578≤X< 0.5996     62 12.4   I□□□□□□□□□□□□□□□□□□□□□
0.5996≤X< 0.6414     46  9.2   I□□□□□□□□□□□□□□□
0.6414≤X< 0.6832     41  8.2   I□□□□□□□□□□□□□□
0.6832≤X< 0.7250     57 11.4   I□□□□□□□□□□□□□□□□□□□
0.7250≤X< 0.7668     32  6.4   I□□□□□□□□□□□
0.7668≤X< 0.8086     39  7.8   I□□□□□□□□□□□□□
0.8086≤X< 0.8504     21  4.2   I□□□□□□□
0.8504≤X< 0.8922     10  2.0   I□□□
0.8922≤X≤ 0.9340      5  1.0   I□□

NUMBER OF MISSING CASES: 1
ONE □ REPRESENTS 3 CASES.

THERE ARE 6 OBJECTS TO BE CLASSIFIED
```

```
HISTOGRAM FOR VARIABLE SPR50

   INTERVAL          FR  PCT   I          +30         +60
0.1810≤X< 0.2207      3  0.6   I□
0.2207≤X< 0.2604      1  0.2   I
0.2604≤X< 0.3001      0  0.0   I
0.3001≤X< 0.3398      4  0.8   I□
0.3398≤X< 0.3795      7  1.4   I□□
0.3795≤X< 0.4192      6  1.2   I□□
0.4192≤X< 0.4589     21  4.2   I□□□□□□□
0.4589≤X< 0.4986     16  3.2   I□□□□□
0.4986≤X< 0.5383     21  4.2   I□□□□□□□
0.5383≤X< 0.5780     25  5.0   I□□□□□□□□
0.5780≤X< 0.6177     36  7.2   I□□□□□□□□□□□□
0.6177≤X< 0.6574     41  8.2   I□□□□□□□□□□□□□□
0.6574≤X< 0.6971     35  7.0   I□□□□□□□□□□□□
0.6971≤X< 0.7368     54 10.8   I□□□□□□□□□□□□□□□□□□
0.7368≤X< 0.7765     46  9.2   I□□□□□□□□□□□□□□□
0.7765≤X< 0.8162     58 11.6   I□□□□□□□□□□□□□□□□□□□
0.8162≤X< 0.8559     47  9.4   I□□□□□□□□□□□□□□□□
0.8559≤X< 0.8956     39  7.8   I□□□□□□□□□□□□□
0.8956≤X< 0.9353     30  6.0   I□□□□□□□□□□
0.9353≤X≤ 0.9750      9  1.8   I□□□

NUMBER OF MISSING CASES: 1
ONE □ REPRESENTS 3 CASES.

THERE ARE 6 OBJECTS TO BE CLASSIFIED
```



```
HISTOGRAM FOR VARIABLE SPJ50

   INTERVAL          FR  PCT   I          +30         +60         +90
0.1270≤X< 0.1686      2  0.4   I□
0.1686≤X< 0.2102      1  0.2   I
0.2102≤X< 0.2518      4  0.8   I□
0.2518≤X< 0.2934      3  0.6   I□
0.2934≤X< 0.3350     10  2.0   I□□□
0.3350≤X< 0.3766      6  1.2   I□□
0.3766≤X< 0.4182     14  2.8   I□□□□□
0.4182≤X< 0.4598     26  5.2   I□□□□□□□□□
0.4598≤X< 0.5014     29  5.8   I□□□□□□□□□□
0.5014≤X< 0.5430     37  7.4   I□□□□□□□□□□□□
0.5430≤X< 0.5846     41  8.2   I□□□□□□□□□□□□□□
0.5846≤X< 0.6262     46  9.2   I□□□□□□□□□□□□□□□
0.6262≤X< 0.6678     35  7.0   I□□□□□□□□□□□□
0.6678≤X< 0.7094     62 12.4   I□□□□□□□□□□□□□□□□□□□□□
0.7094≤X< 0.7510     40  8.0   I□□□□□□□□□□□□□
0.7510≤X< 0.7926     55 11.0   I□□□□□□□□□□□□□□□□□□
0.7926≤X< 0.8342     31  6.2   I□□□□□□□□□□
0.8342≤X< 0.8758     29  5.8   I□□□□□□□□□□
0.8758≤X< 0.9174     22  4.4   I□□□□□□□
0.9174≤X≤ 0.9590      7  1.4   I□□

NUMBER OF MISSING CASES: 0
ONE □ REPRESENTS 3 CASES.

THERE ARE 6 OBJECTS TO BE CLASSIFIED
```

```
HISTOGRAM FOR VARIABLE SPERM

   INTERVAL          FR  PCT   I          +30         +60
0.1660≤X< 0.2026      3  0.6   I□
0.2026≤X< 0.2392     10  2.0   I□□□
0.2392≤X< 0.2758     13  2.6   I□□□□
0.2758≤X< 0.3124     18  3.6   I□□□□□□
0.3124≤X< 0.3490     30  6.0   I□□□□□□□□□□
0.3490≤X< 0.3856     21  4.2   I□□□□□□□
0.3856≤X< 0.4222     26  5.2   I□□□□□□□□□
0.4222≤X< 0.4588     36  7.2   I□□□□□□□□□□□□
0.4588≤X< 0.4954     38  7.6   I□□□□□□□□□□□□□
0.4954≤X< 0.5320     51 10.2   I□□□□□□□□□□□□□□□□□
0.5320≤X< 0.5686     41  8.2   I□□□□□□□□□□□□□□
0.5686≤X< 0.6052     48  9.6   I□□□□□□□□□□□□□□□□
0.6052≤X< 0.6418     30  6.0   I□□□□□□□□□□
0.6418≤X< 0.6784     31  6.2   I□□□□□□□□□□
0.6784≤X< 0.7150     30  6.0   I□□□□□□□□□□
0.7150≤X< 0.7516     26  5.2   I□□□□□□□□□
0.7516≤X< 0.7882     13  2.6   I□□□□
0.7882≤X< 0.8248     21  4.2   I□□□□□□□
0.8248≤X< 0.8614     11  2.2   I□□□□
0.8614≤X≤ 0.8980      3  0.6   I□

NUMBER OF MISSING CASES: 0
ONE □ REPRESENTS 3 CASES.

THERTE ARE 6 OBJECTS TO BE CLASSIFIED
```

FILMED

-8